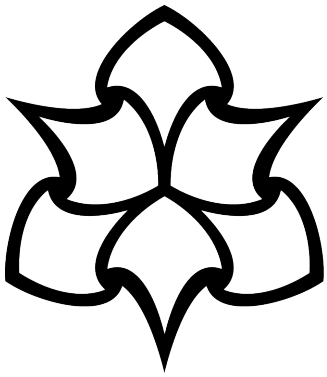

Mining the Context of Citations in Scientific Publications

Saeed-Ul Hassan[1[0000-0002-6509-9190]], Sehrish Iqbal[1[0000-0003-1956-1572]], Mubashir Imran[2[0000-0003-4721-499X]], Naif Radi Aljohani[3[0000-0001-9153-1293]] and Raheel Nawaz[4[0000-0001-9588-0052]]

[1] Information Technology University, Ferozepur Road, Lahore, Pakistan
[2] The University of Queensland, St Lucia QLD 4072, Australia
[3] King Abdulaziz University, Al-Malae'b St, Jeddah, Saudi Arabia
[4] Manchester Metropolitan University, Manchester M15 6BH, United Kingdom
saeed-ul-hassan@itu.edu.pk

Abstract. Recent advancements in information retrieval systems significantly rely on the context-based features and semantic matching techniques to provide relevant information to users from ever-growing digital libraries. Scientific communities seek to understand the implications of research, its importance and its applicability for future research directions. To mine this information, absolute citations merely fail to measure the importance of scientific literature, as a citation may have a specific context in full text. Thus, a comprehensive contextual understanding of cited references is necessary. For this purpose, numerous techniques have been proposed that tap the power of artificial intelligence models to detect important or incidental (non-important) citations in full text scholarly publications. In this paper, we compare and build upon on four state-of-the-art models that detect important citations using 450 manually annotated citations by experts - randomly selected from 20,527 papers from the Association for Computational Linguistics corpus. Of the total 64 unique features proposed by the four selected state-of-the-art models, the top 29 were chosen using the Extra-Trees classifier. These were then fed it to our supervised machine learning based models: Random Forest (RF) and Support Vector Machine. The RF model outperforms existing selected systems by more than 10%, with 89% precision-recall curve. Finally, we qualitatively assessed important and non-important citations by employing and self-organizing maps. Overall, our research work supports information retrieval algorithms that detect and fetch scientific articles on the basis of both qualitative and quantitative indices in scholarly big data.

Keywords: Citation Context Analysis, Influential Citations, Machine Learning, Self-Organizing Maps

1 Introduction

The measure of impact generated by scientific literature is often accounted by the citations it received [1], Based on absolute citation counts, numerous bibliometric measures (such as H-Index, G-index, and SNIP etc.) have been introduced over the

years. Whilst, such measures reflect upon different quantitative aspects of a scholarly literature impacts. It has been disputed whether these measures also provide insights regarding the impact, the cited work had within a scientific literature? To address this issue, it is suggested to look into the qualitative aspects of a citation.

Moravcsik and Murugesan [2], deconstructed citations identified in scientific literature into four dimensions; a) Conceptual Use or Operational Use (was some theory used as support or some technical method/equation adopted); b) Evolutionary or Juxtapositional (cited work is base or alternate); c) Organic or Perfunctory (does the citing work explain certain point or is it just a general acknowledgement); and d) Confirmative vs. Negational (does the citation claim correctness or dispute). The study shows that a major portion (40% cited articles) were general acknowledgements. This reinforces the importance of citation context. Numerous studies discuss the issue of identifying the importance of citations using supervised machine learning techniques applying contextual and quantitative features [3-5]. The algorithms and techniques to approach a certain research problem, as well as the writing style of the author [6], contribute greatly in making an article influential. The number of citations received by scientific literature often accounts for their quantitative impact, but not all citations can be considered equal. Thus, to understand the influence of cited work in citing work, categorizing citations into levels of importance and incidental class is essential.

In this paper, we aim to address the problem of classifying cited work as important or incidental. The followings are the contributions of our paper: At first, we compare four state-of-the-art citation classification techniques using the dataset downloaded from ACL Anthology corpus[1] of 20,527 publications. Further, we present 29 features that outperform existing state-of-the-art techniques by extracting top features from all four selected techniques. Finally, we qualitatively analyze the distinction between important and non-important citations by employing self-organizing maps.

2 Literature Review

A citation context is essentially the text surrounding reference markers. Conventionally, citation analysis has been used to measure the quality of articles in scholarly literature, hence the tracking of citations plays a vital role. It has been argued by Valenzuela [4] that all citations are not equally important, therefore, classification is needed to distinguish the important ones from the unimportant. Xu et al. [7] proposed a citation classification technique by using three classes: functional, ambiguous and perfunctory. They used heterogeneous set of features for classification i.e. cue patterns, positional features, network-based features and structural features to measure the relationship between the author and the article. According to their results combining these features is a challenging task but individually their features perform

[1] http://www.aclweb.org/anthology/

well. Citation analysis has been used widely to detect scientific collaboration patterns, observing knowledge graphs, impact measurement etc.

Cohan and Goharian [8] addressed the problem of inaccurate citation context extraction, they presented a framework for the automatic summary of research articles by using the context of citations. The framework is consisted of three parts: a) to find context of citation b) identify features of the citation context c) and generating a summary of citation contexts. A new study of multiple in-text references (MIR) with respect to their position in the article and syntactic context has been proposed by Bornmann et al. [9]. They used a dataset of 80,000 research articles for analyzing two characteristics: a) the position of the MIR in rhetorical structure of article and b) the total number of references in context that make a MIR. Presence of MIR implies the presence of features i.e. topic, keywords and methods common to work cited in aggregated of in-text reference.

3 Data and Methodology

We obtained data from the data corpses of Association for Computational Linguistics (ACL)[1] containing 20,527 articles that are publicly available. These articles contained 106,509 citations. Within these citations 450 unique citations were randomly chosen and labeled as important or unimportant/incidental by a group of field experts [4]. This labeling was further authenticated by a group of experts in the field of computational linguistics. Out of these 14.6% of the citations were considered important by the experts while the rest were marked as unimportant.

3.1 Citation Classification Models

In this section, we define the data extraction, select state-of-the-art supervised machine learning model [1, 3-5] and construction of our supervised and unsupervised models.

Teufel Model. Teufel et al. [3] recommended a technique to categorize the citation function automatically by extracting sets of features i.e. (shallow and linguistically inspired features, part-of-speech-based recognition and finite grammar using string). Each feature is classified into four categories, namely; weakness, comparison, sentiments and neutral. These features are further categorized as weak, positive or neutral. Using the supervised classification model, an accuracy of 83% was attained.

Amjad Model. Amjad et al. [1] used Teufel et al. [3] work and proposed different context-level and polarity–level features for; a) reference tagging b) reference grouping and c) non-syntactic reference removal and polity. For categorizing citations SVM (kernel=linear, c=1.0) with 10-folds cross-validation was used for context identification and attained a precision of 92% on a recall of 76.4%. For citation purpose classifications 70.5% accuracy has been attained.

Valenzuela Model. Valenzuela et al. [4] argued that we cannot categorize citations as being of equal importance. Hence, they proposed a citation categorization mechanism

into important and non-important class. For this purpose, they mined 12 new features mostly related to the nature of the reference and the section in which it is cited. They constructed a supervised classification model with SVM (kernel=RBF) and RF. Both classifiers attained an encouraging area under the curve (AUC) of 80%.

Hassan Model. Hassan et al. [5] extended Valenzuela's et al. [4] work and presented 13 features categorized into three groups; context-based features; cue word-based features; and textual features. They constructed a model with five classifiers, namely Random Forest, SVM, KNN, Decision Tree and Naïve Bayes. RF outperforms other classifiers, with an AUC of 91%.

Hassan_29 Model. By combining the features of all models [1, 3-5], there are a total 64 features. To extract the best features, we employed the Extra-Trees classifier proposed by Geurts et al. [10], that divide the complete selection of data at each step and randomly picks a decision boundary. Finally, we elected all 29 features that had an Extra-Trees classifier score greater than 1%. We named our machine learning model as 'Hassan_29'.

Deployed Unsupervised Models. To obtain a better understanding of the data we devised unsupervised approaches. This helps us to better visualize how our data set and each distinct feature point contained within it behaves. The unsupervised approach used Self Organizing Maps (SOM). This study employed a 10x10 (neurons) SOM on Hassan_29 to observe the behavior of cited literature.

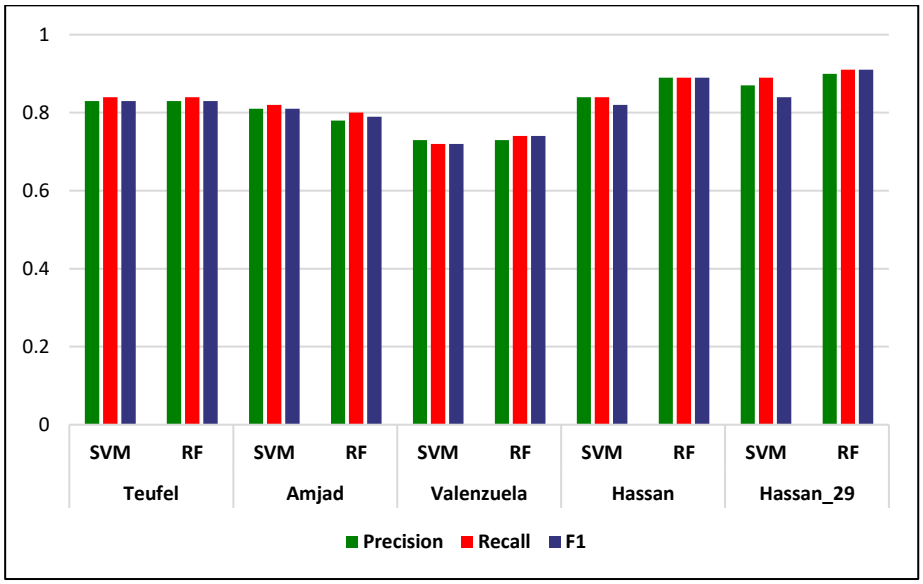

Fig. 1: Precision, Recall and F1 score computed by SVM and RF on four state of the art models and newly proposed Hassan_29 model.

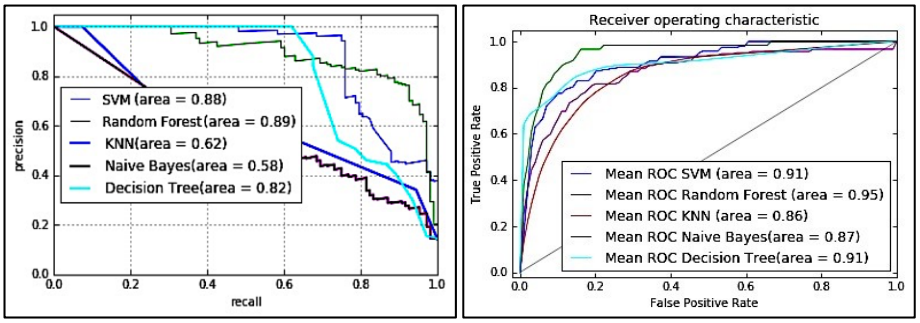

Fig. 2: (left) Precision recall curve and (right) ROC curves for Hassan_29 model using SVM, RF, KNN, Naïve Bayes and Decision Tree classifiers.

4 Experimental Settings and Results

We compared the performance of features extracted using four state-of-the-art techniques and 'Hassan_29' features by training models on SVM [11] and RF [12]. The SVM finds the best boundaries of the outputs by converting data using a specific kernel. Here, we applied a non-linear Radial Basis Kernel (RBF) for transformation [13]. The RBF function is provided in Eq. 1.

$$k(x,z) = \mathrm{e}^{-\gamma} \parallel x - z \parallel 2, \gamma > 0 \tag{1}$$

Here $\mathrm{e}^{-\gamma}$ is a constant, while x and z denote vectors in some feature space. Random Forest is an algorithm that, as the name suggests, creates a forest of classification trees and splits the feature nodes randomly. We calculated precision, recall and F1-score to compare the performance of each model on the same dataset shown in Fig. 1. To divide the data into training and testing samples three-fold cross validation technique was used. Fig. 1 shows that our set of features i.e. 'Hassan_29' outperforms all other models, having better precision over a high recall, with an f-measure reaching 0.91 for the RF classifier. Note, on this set of data, RF shows better performance as compared to SVM as indicated by the graph. The main reason behind better performance of RF is that the features of these models consist of a mixture of continuous and numeric features, as well as outliers. In such cases, Random Forest performs well.

Drilling down further, we evaluate the effectives of Hassan_29 model by deployment precision recall and ROC curves (see Fig 2). Findings suggest that, RF classifier outperforms other classifiers with PR curve of 89% and ROC 95%. The Naïve Bayes classifier performed worst, which could be because the model fails to learn interaction among the features as our dataset is relatively small.

For better data visualization and a qualitative understanding of features, we apply SOM to reduce the data dimension to 2D.

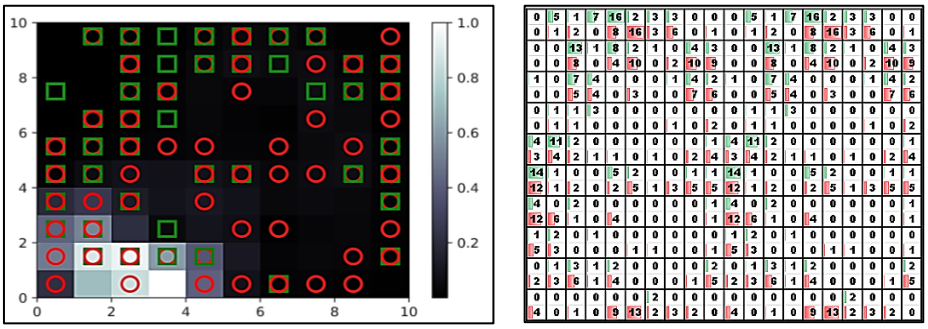

Fig. 3: (left) Heat map of SOM classifying important (green box) /incidental (red circle) citations and (right) density of citations mapped on each neuron. Top row (green) of each cell represents important while bottom (red) row represents incidental.

Fig. 3(left) represents a heat map of SOM neurons. The background represents the average distance map of the weight, where lighter color (white) represents greater distance (lesser weight), while dark color (black) represents lesser distance (greater weight). The green and red marker represents each class from the dataset and their position corresponds to the neuron on which they are classified. Fig. 3(right) represents the mapping of each citation to a certain neuron. Positive classes form independent tight large clusters, with many neurons (e.g. at (1,4), (2,4)), however non-important class performed better creating large independent cluster at adjacent neurons. This makes it is easier to identify non-important citations as compared to important.

5 Concluding Remarks

We have compared and build upon four state-of-the-art models that address the problem of classifying citations into important and non-important groups. We have shown that our machine-learning model, with top 29 features, outperforms all existing state-of-the-art models. In addition, we found that qualitative assessment helps better understand the feature set being examined. A potential limitation of this study is the adoption of the definitions that as such came with the dataset [4, 14]. In future studies, other definitions and features could be explored, such as stylometric features from full-text [15].

Overall, our proposed technique contributes to Bibliometric Enhanced Information Retrieval system by increasing query search capabilities of search engines. Moreover, citation classification can be used to qualitatively measure the impact of publications in our growing scholarly big data and in the behavioral analysis of scientific domains. Finally, this study can help to improve citation-based full text summarization techniques.

Note that the data and the code used in this paper can be downloaded from the following URL: https://github.com/slab-itu/citation_context_icadl_2018.

References

1. Abu-Jbara, A., Ezra, J., Radev, D. Purpose and polarity of citation: Towards nlp-based bibliometrics. In: Proceedings of the conference of the North American chapter of the association for computational linguistics: Human language technologies pp. 596-606 (2013).
2. Moravcsik, M. J., Murugesan, P. Some results on the function and quality of citations. Social studies of science, 5(1), 86-92. (1975).
3. Teufel, S., Siddharthan, A., & Tidhar, D. Automatic classification of citation function. In: Proceedings of the conference on empirical methods in natural language processing. pp. 103-110, Association for Computational Linguistics (2006).
4. Valenzuela, M., Ha, V., Etzioni, O. Identifying Meaningful Citations. In: AAAI Workshop: Scholarly Big Data (2015).
5. Hassan, S. U., Safder, I., Akram, A., and Kamiran, F.: A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis. Scientometrics, 116(2), 973-996 (2018).
6. Hassan, S. U., Imran, M., Iftikhar, T., Safder, I., Shabbir, M. Deep Stylometry and Lexical & Syntactic Features Based Author Attribution on PLoS Digital Repository. In: International Conference on Asian Digital Libraries, pp. 119-127, Springer (2017).
7. Zhu, X., Turney, P., Lemire, D., Vellino, A. Measuring academic influence: Not all citations are equal. Journal of the Association for Information Science and Technology, 66(2), 408-427(2015).
8. Cohan, A., Goharian, N., Scientific document summarization via citation contextualization and scientific discourse, International Journal on Digital Libraries, 1-17, (2017).
9. Bornmann, L., Haunschild, R., Hug, S. E., Visualizing the context of citations referencing papers published by Eugene Garfield: A new type of keyword co-occurrence analysis. Scientometrics, 114(2), 427-437, (2018).
10. Geurts, P., Ernst, D., Wehenkel, L. Extremely randomized trees. Machine learning, 63(1), 3-42, (2006).
11. Auria, L., Moro, R. A. Support vector machines (SVM) as a technique for solvency analysis. Technical report, Deutsche Bundesbank, Hannover; German Institute for Economic Research, Berlin. (2008).
12. Breiman, L. Random forests. Machine learning, 45(1), 5-32, (2001).
13. Cao, H., Naito, T., & Ninomiya, Y. Approximate RBF kernel SVM and its applications in pedestrian classification. In The 1st International Workshop on Machine Learning for Vision-based Motion Analysis-MLVMA (2008).
14. Hassan, S. U., Akram, A., Haddawy, P. Identifying Important Citations Using Contextual Information from Full Text. In Digital Libraries (JCDL) ACM/IEEE Joint Conference on (pp. 1-8). IEEE (2017).
15. Hassan, S. U., Imran, M., Iftikhar, T., Safder, I., and Shabbir, M.: Deep Stylometry and Lexical & Syntactic Features Based Author Attribution on PLoS Digital Repository. In: International Conference on Asian Digital Libraries, pp. 119-127. Springer, Cham (2017).